Guerra *et al.*
Journal on Image and Video Processing (2026) 2026:16
https://doi.org/10.1186/s13640-026-00701-x

Journal on Image and Video Processing

RESEARCH Open Access

A new dataset and pipeline for ICAO compliance check of head coverings: an inclusive religious-aware approach

Carla Guerra[1*], Diogo Nunes[1], David Carreira[1], João Marcos[1] and Nuno Gonçalves[1]

*Correspondence:
carla.guerra@isr.uc.pt

[1] Institute of Systems and Robotics, University of Coimbra, Coimbra, Portugal

Abstract

The automatic verification of the compliance of Machine-Readable Travel Documents' (MRTD) portraits is nowadays of big interest for issuing authorities or application providers. Despite the increase of performance and effectiveness of automatic services, there is still much room for the improvement of such systems. Specific particularities related to cultural or religious aspects may increase inequalities or social tension between different people. Additionally, regulations and standards allow portraits a few differences related to these aspects, yet current automatic systems to verify their compliance generally do not take into account these particularities. This paper then contributes with a dataset and an algorithm that automatically verifies the compliance with the ICAO requirements related to the use of head coverings on facial images used on MRTDs. All the methods found in the literature ignore that some coverings might be accepted because of religious or cultural reasons, and mainly only look for the presence of hats or caps. Our approach specifically includes the religious cases and distinguishes the head coverings that might be considered compliant, depending on the national rules. We built a dataset composed of 6759 facial images of 910 identities to accommodate these types of head coverings. That data was used to fine-tune and train a new classification model derived from the YOLOv8 framework and we achieved state-of-the-art results with an overall accuracy on the created data of 91.05% and HTER of 8.53%.

Keywords: Facial images, ICAO, ISO/IEC 19794-5, Head covering compliance, Religious head coverings

1 Introduction

Photographs used in ID documents, particularly in Machine-Readable Travel Documents (MRTD), usually called portraits, must comply with certain requirements that guarantee standardization, in addition to allowing the person represented in the portrait to be properly identified through this image. Compliance with these requirements is based on quality metrics that measure, for instance, the framing of the head in the photograph, the contrast with an homogeneous background, the restriction on the use of sunglasses or glasses whose lenses or frames partially or completely cover the eyes, among many others.

© The Author(s) 2026. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

Springer

Three of the most relevant and extended public documents related to quality assessment in biometrics are the ISO/IEC 19794-5 standard [1, 2], Doc 9303 [3] and Technical Report on Portrait Quality [4] both released by the International Civil Aviation Organization (ICAO), based on that standard. These documents are actually a series of guidelines for the acquisition of high-quality images, i.e., portrait-like images, for their inclusion on MRTD, such as passports and ID cards. These guidelines are based on the typical impact that certain features like blur, occlusions, and resolution have in the quality of facial images and are intended to preserve the performance of Facial Recognition Systems (FRS). However, these reports do not specify the method to measure each one of the features. In order to implement their recommended guidelines it becomes necessary to develop algorithms to automatically verify the compliance with the requirements. This can be verified using image processing techniques, combined with geometric measurements of elements detected in the photograph or with more sophisticated methods including deep learning.

In this work, we focus on a specific requirement related to the use of head coverings in the photograph. Head coverings should not be accepted except in circumstances specifically approved by each national issuing authority for the MRTDs [5]. Such circumstances may be religious, medical or cultural.

Regulations then allow the presence in portraits of head coverings for religious circumstances, but adds that, in such situations, veils, scarves or head coverings "shall be firm fitting and of a plain uniform color with no pattern and no visible perforations and the region between hair lines, both forwards of the ears and chin including cheeks, mouth, eyes, and eyebrows shall be visible without any distortion or shadows" [5].

In this context, to the best of our knowledge, current disclosed solutions for the automatic verification of the ICAO compliance of head coverings completely ignore religious circumstances, and look only for the presence of something covering the head.

The main contribution of this paper are, thus, twofold:

- We built our own dataset for the training and testing of the algorithm, given the fact that none of the public datasets found on the literature consider the relevant particularities needed to verify ICAO compliance when in the presence of religious coverings. The results obtained with the proposed method outperformed all the state of the art methods.[1]
- We propose a pipeline that automatically verifies the compliance with the requirements related to the use of head coverings, specifically considering religious cases.

To the best of our knowledge, both contributions, in the form of the dataset and method for the compliance check of this requirement, are the first religious-aware approaches to date. For simplicity, in the remainder of this paper we will only reference both of them as head coverings.

This article is organized as follows: first we give an overview of the literature work related to the verification of ICAO requirements regarding head coverings (Sect. 2). Then we present the dataset we built for this purpose in Sect. 3 and we propose our

[1] Check the public repository at https://github.com/visteam-isr-uc/DFIC-HC

Table 1 Published results on the FVC-onGoing platform regarding the 'Hat/Cap' ICAO requirement

	ICAONet	id3	BioPass Face	BioTest
EER	**5.7%**	6.8%	9.8%	16.5%

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

method to automatically verify the compliance with the head coverings requirements in Sect. 4. We show and discuss the results in Sect. 5, and we finish with the conclusions and future work in Sect. 6.

2 Related work

The University of Bologna's Biolab group played a significant role in popularizing methods adhering to the ISO/IEC 19794-5 standard. In 2009, they introduced the Biolab-ICAO framework [6], a novel and independent benchmark tool for assessing the compliance of face images within ICAO requirements that addresses limitations regarding the non-existence of images that failed certain requirements, namely the presence of hats or caps. Subsequently, in 2012 the benchmark underwent further refinement, compiling images from public datasets, namely AR [7], FRGC [8], PUT [9] and private samples, which resulted in a global collection of 5588 images from 601 identities divided into an official benchmark database (4868 pictures) and testing protocol (720 pictures). Recently, Biolab is also making an effort to augment the datasets available through high-quality synthetic images of faces [10].

Biolab has also presented the first documented proposal that describes how to evaluate 23 scene requirements—the BiolabSDK [11]. This work established a major starting point, however, by today's standards and increasing necessity of security, the use of classical computer vision algorithms with human engineered features is a drawback. The availability of Deep Learning-based methods, as can be seen in the following sections, proves to be a better promoter of accurate predictions and generalization, ultimately leading to less biased and more inclusive approaches.

As of today, the Biolab-ICAO framework is used to evaluate algorithms via an online public competition called Face Image ISO Compliance Verification (FICV), hosted at the FVC-onGoing website [12]. The FICV is considered the standard evaluation tool for ISO/IEC 19794-5 standard and is used by the most relevant works presented in the literature or commercial products. 23 scene requirements are evaluated individually and the results are reported in terms of: EER, FMR_{100}, $Zero_{FMR}$, $Zero_{FNMR}$, Rejection, Impostor and Genuine score distributions and the FMR, FNMR and DET curves.

2.1 Head coverings ICAO requirement

Out of the 23 requirements, we are focusing, with particular attention, on the 'Hat/Cap' requirement, a less representative version of what we defined in this work as 'Head Coverings', which accommodates many more options, including religious scenarios that should not be simply treated and processed as hats or caps and that may be compliant if some particular restrictions are met and depending on each national issuer authority.

To date, there are four published algorithms in the FVC-onGoing platform regarding the 'Hat/Cap' requirement: BioTest, BioPass Face, id3, and ICAONet [13]. Their results

are directly extracted from the benchmark's published results and are summarized on Table 1.

Three out of the four algorithms are proprietary solutions owned by private companies, therefore, there is no detailed explanation about their methods. The remainder one is ICAONet, which makes use of the significant advancements observed in Deep Learning over the past decade. It leverages a Multitask Learning (MTL) approach that automatically evaluates the 23 requirements at the same time, achieving state of the art results in nine of them. ICAOnet utilizes a single neural network composed of three main components: an encoder, a decoder and a dense branch. The two first components constitute an auto-encoder. The encoder takes the images and produces a feature embedding that is shared with the decoder for unsupervised image reconstruction, and with the dense branch for multi-label classification. This work proposes a robust and efficient approach to solve all the requirements, however, due to its Deep Learning-based nature, its bottleneck will be the quality and quantity of the available data. ICAOnet requires a balanced dataset to ensure good quality results, so that during the training phase, the network does not learn toward a specific pattern and significantly disregards others. Moreover, this work also did not take into account the diverse possibilities of the 'hat/cap' scenario.

Another very relevant work (that is not already published in the FVC-onGoing platform) is the "Open Source Face Image Quality" (OFIQ), an open-source software for the quality assessment of facial images used in face recognition [14]. It serves as the reference implementation of the algorithms outlined in the recent revision of ISO/IEC 29794-5:2025 and is intended for use in various biometric applications, with a particular focus on border control settings. Nonetheless and as before, the algorithm to verify the presence of head coverings disregards religious circumstances where certain types of head coverings might be considered compliant. The algorithm takes as input a face parsing segmentation map and then computes the ratio between pixels mapped as head covering or clothing and the total number of pixels. The method used to create the segmentation map does not distinguish between compliant or non-compliant coverings, it only detects the presence of some covering or not.

There are several other studies on how to solve certain requirements, however, they are not within the scope of this paper. Hence, we also highlight the recent work by Hernandez-Ortega *et al.* [15], who proposed FaceQvec for evaluating the conformity of facial images within the same 23 requirements defined by BioLab, plus two regarding white-noise estimation and facial expressions. It employs a mix of classical approaches and pre-trained Convolutional Neural Networks (CNN) to evaluate them. Regarding the head coverings in particular, the algorithm searches for unnatural colored pixels in the upper forehead region, making the 'Hats/Caps' distinction without taking into account possible religious coverings. Furthermore, the authors did not undergo the BioLab benchmark and do not show their own results regarding this requirement in particular due to the lack of negative samples for model development and testing.

After the literature review we then concluded that there is already some work on the automatic verification of the head coverings' requirements, but there is still the need to extend these methods to consider more than hats and caps to other types of head coverings. Furthermore, there is a need to implement algorithms for a broad choice of head coverings, particularly extending the analysis to religious, medical or cultural

circumstances. We presented preliminary work towards this goal already at the 21st International Conference of the Biometrics Special Interest Group (BIOSIG23) [16], making this paper an extended version of that work.

2.2 Regulations and standards

For the compliance of face portraits for MRTD documents, there are several important documents to take into account. The International Civil Aviation Organization (ICAO) establishes the standards for MRTDs since the first version of the technical document Doc 9303 published in 1980, whose title is "A Passport with Machine Readable Capability". This document evolved to a second edition with the addition of biometrics in part 2 in 2006. The last edition of the ICAO's Doc 9303 was in 2021, with a detailed description of the requirements as well as recommendations, including the printing process and its scanning, for instance so that MRTD can be used in Automated Border Control (ABC) gates [3].

As for the attributes of the portraits, and relating directly to the use of head coverings, the first generation of biometric data interchange format (published between 2005 and 2007) and the second generation published from 2011 onwards, both of the ISO/IEC 19794 standard series, established the specific rules so that a portrait can be compliant, particularly specifying the conditions for the use of head coverings in the ISO/IEC 19794-5-Face Image Data in both versions of 2005 [17] and 2011 [1].

Additionally, as the requirements have evolved and as it was expected to be necessary in the near future to expand the specifications, it was created a third generation of the series in the standard of ISO/IEC 39794:2019, particularly the part 5 for the specification of the compliance of using head coverings [18]. This generation series (ISO/IEC 39794:2019) is incompatible with ISO/IEC 19794:2005 and ISO/IEC 19794:2011 but it is expected that the latter ones will remain valid as long as they are needed since most of the current MRTDs still follow the first or second version of the standard series. Concerning the transition set by ICAO, border inspection systems must support the ISO/IEC 39794-5 format from 1 January 2026; issuing authorities may use either format between 2026 and 2030; and from 1 January 2030 all newly issued eMRTDs must use ISO/IEC 39794-5, with ISO/IEC 19794 fully deprecated by 2040.

Even if the face portrait specifications about what is compliant and what it is not compliant are specified in the previous described standards of 2005 and 2011, in 2018 ICAO and the joint standardization group ISO/IEC JTC1 SC17 WG3 have published a document named "Portrait Quality (Reference Facial Images for MRTD)" that rearranges, consolidates, enriches and improves the ISO/IEC 19794-5:2005, ISO/IEC 19794-5:2011 and Doc 9303, eventually published as a ICAO's Technical Report [4].

Furthermore, the face image quality assessment has been studied in order to automatically measure the face quality. Standardization attempts, to harmonize solutions also reported by the industry and with strongly different concepts of what is ICAO compliant or not, led to the publication of the Technical Report ISO/IEC 29794-5, Information technology—Biometric sample quality—Part 5: Face image data [19]. More recently, the working group ISO/IEC JTC 1/SC 37/WG 3 replaced this Technical Report by the standard ISO/IEC 29794-5, published as an International Standard in April 2025 [2]. This standard, which is aligned with both the ISO/IEC 19794-5:2011 and ISO/IEC

39794-5:2019, brings a unified quality score and discusses capture-related and subject-related quality components of face images.

Finally, the need to harmonize the Biometrics vocabulary in standards, reports and other documents, motivated the publication of the standard ISO/IEC 2382-37:2022, which is a very useful tool to increase harmonization [20].

3 Diverse face images Coimbra—head coverings dataset (DFIC-HC)

To overcome the lack of publicly available datasets to develop algorithms to validate the compliance with ICAO requirements and, in particular, the 'Head Coverings' one, we built a large-scale facial image dataset called Diverse Face Images Coimbra—Head Coverings (DFIC-HC).

A smaller version of this dataset (as well as an older version of the proposed method to verify the ICAO compliance regarding head coverings) was already presented in the proceedings of the 21st International Conference of the Biometrics Special Interest Group (BIOSIG23) [16]. This paper brings an extended version of the dataset, with almost double the images, manually revised labels and an extensive evaluation of the proposed method on different conditions.

The current version of the dataset contains a total of 6759 high-quality facial images, taken from 910 identities gathered across volunteers of different ages, genders and origins. The images were collected in controlled conditions, using many different head accessories, including religious options that could be considered compliant or not.

The DFIC-HC is a subset of a dataset of higher dimensions (DFIC) with around 24 images per identity, with more than 1050 identities as a whole. This dataset is currently being finalized and labeled, to be submitted for publication soon. The DFIC-HC dataset will be publicly published with the final version of this article.

3.1 Participants

The current publicly available datasets used in this context do not present enough age and ethnic diversity, as shown in Table 2. Most of these datasets do not even specify their age and origin distributions. In smaller dimension datasets such as AR [7] and PUT [9] we did the annotation manually. Our manual annotations may have a certain noise because it is hard to infer age and origin solely by looking at pictures of that person. However, this process is enough to provide a general overview of the datasets demographics. Due to the high number of images in CelebAMask-HQ dataset [21], we did not annotate it. Some annotations regarding gender and age were already specified in the literature, distinguishing only between 'Young' and 'Not Young' in terms of age. By inspecting some of the images and corresponding annotation we suggest that the threshold between 'Young' and 'Not Young' labels is within the interval from 30 to 40 years old. FRGC [8] provides all the information, but it considers much less age diversity. Overall, there is not much balance across groups.

There is some work that tries to mitigate this lack of balance through the selection of invariant features that do not depend so much on large and diverse amounts of data, such as with age-related diversity [22], but its application in the verification of many of the ICAO requirements is still hard.

Table 2 Overview regarding gender, origin and age distributions over the existing datasets used in research related to the verification of ICAO requirements concerning head coverings

Dataset	#Subjects	#Images	Gender		Origin			Age
			Male	Female	Asian	Caucasian	Other	Age intervals (yo)
PUT	100	9,971	89%	11%	0%	100%	0%	[0–18[: 0%; [18–40[: 97%; [40+]: 3%
AR	136	3,315	56%	44%	3%	97%	0%	[0–18[: 1%; [18–40[: 74%; [40+]: 26%
CelebAMask-HQ	–	30,000	37%	63%	–			[Young]: 78%; [Not Young]: 22%
FRGC	466	40,000	57%	43%	22%	68%	10%	[18–22]: 65%; [23–27]: 11%; [28+]: 17%
DFIC-HC	910	6,759	49%	51%	7%	79%	14%	[0–20]: 34%;]20–35]: 40%;]35–50]: 13%;]50–65]: 7%;]65+]: 6%
DFIC-HC (B)	337	2,505	46%	54%	26%	41%	33%	[0–20]: 22%;]20–35]: 24%;]35–50]: 21%;]50–65]: 16%;]65+]: 17%

Some of the entries were not specified by the authors—in smaller datasets such as AR and PUT, we manually annotated them; in the high dimension CelebAMask-HQ dataset we marked the entries not annotated with '-'. 'DFIC-HC (B)' refers to a balanced subset of DFIC-HC

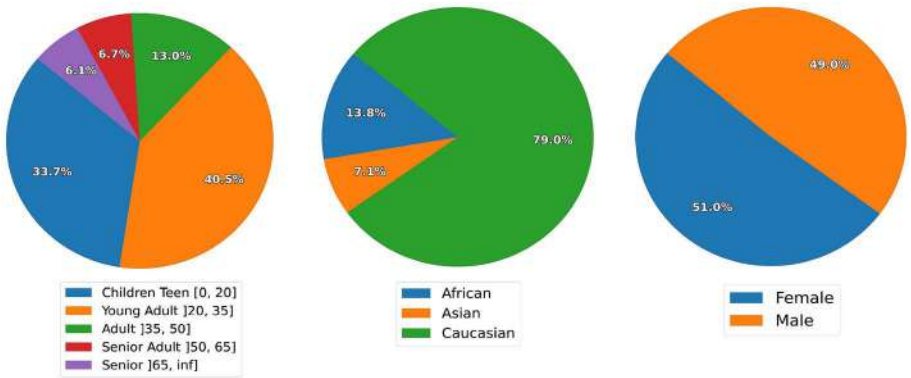


Fig. 1 Percentage of each class in the different DFIC-HC dataset categories. (Left) relates to the age; (middle) describes the different ethnic origins' distribution; and (right) is the gender distribution

To overcome this shortcoming, we collected facial images with a special care about diversity, including participants of different ages, from toddlers to seniors and ethnic backgrounds. We classified the identities regarding age, gender and ethnic origin. As for the representation of children and seniors we verified that currently available datasets have a noticeable gap for these specific groups, under-representing them. Therefore, we made additional efforts to include participants from these demographic groups.

Figure 1 shows the demographic distribution of all the identities collected. Regarding the gender attribute we considered only two classes—male (M) or female (F). The age attribute includes 5 groups—children/teen (age between 0 and 20 years), young adult (age between 21 and 35 years), adult (age between 36 and 50 years), senior adult (age between 51 and 65 years) and senior (age above 65 years). These groups were defined

following the National Institute of Standards and Technology (NIST) standards, particularly for the FRTE benchmark to Demographic studies [23]. The ethnic origin of the participants was divided into 3 groups—'Caucasian', 'African' and 'Asian'—although one can fine-grain it if needed because, during the dataset acquisition, the country of origin of each identity was registered. Some datasets used in this context use the term 'White' to refer to the group we called 'Caucasians'—we tried to choose the term that sounded more inclusive and respectful to us, although cultural differences may require different terms.

The new dataset DFIC-HC significantly improves the balance level across the three attributes of gender, age and origin when compared with any other dataset found in the literature that has been used for ICAO compliance evaluation.

3.2 Images collected

The images were acquired in single sessions for each participant. The participants were summoned between several groups, including students, researchers, professors, staff and their families of the University of Coimbra and also a secondary school in Coimbra, foreign groups living in Coimbra (mainly African), African festivals occurring in the city, a shopping mall and a day center for the elderly in Coimbra, international conferences in Lisbon, specific neighborhoods in Lisbon and in the Center-East of India, where a researcher of the group was temporarily working.

All the pictures were voluntarily taken, that is, if a participant was not comfortable with a specific scenario, they would not be forced to take that picture. That being said, each volunteer was proposed 7 scenarios based on ICAO's standards:

- 2 with no head coverings at all;
- 3 with non-compliant head coverings (hat, caps, ribbons, etc.);
- 1 with religious coverings considered compliant;
- 1 with non-compliant religious coverings.

An example of the complete list of pictures taken of each participant can be seen in Fig. 2.

The image capture setup varied slightly depending on the many places where the acquisition sessions took place. Most of the pictures were collected in indoor rooms with controlled artificial lighting, including two softboxes and a small portable light to purposely create shadows. A big white poster right behind each subject was also used, mainly to guarantee an homogeneous background when needed. Only a small portion of the pictures were taken with uncontrolled lighting conditions. The subjects were positioned between 1 to 1.5 m from the camera that was kept in a tripod. Two different cameras were used—a Canon EOS 350D with a 18–55 mm zoom lens and a Panasonic Lumix S5 with a 20–60 mm zoom lens.

In the end, the dataset was composed of 6759 images:

- 2897 images that, depending on the MRTD issuer and the national rules, may be considered compliant;



Fig. 2 Samples of the pictures taken by each identity on the dataset. Top 3 are compliant regarding the requirements related to head coverings, the bottom 4 are non-compliant

- 3862 images that are non-compliant with at least one of the ICAO requirements regarding head coverings.

3.3 Types of head coverings

ICAO Doc 9303 and ISO/IEC 39794-5 state that if head coverings are allowed, they shall be firm fitting and of a plain uniform color with no pattern or no visible perforations and the region between hair lines, both forwards of the ears and chin including cheeks, mouth, eyes, and eyebrows shall be visible without any distortion or shadows [3, 4, 18].

To guarantee a considerable variety of head coverings, a collection of many different items was provided to each participant, from religious options (e.g., hijabs, gutras) to non-religious ones (e.g., caps, hats, beanies). Table 3 summarizes the complete list of items and the number of images with each one. Figure 3 shows a collection of those items. To note that when the participant had his own head covering, we prioritized its use to increase the variety. If a certain head covering does not fit any of the categories in Table 3, it is classified as 'Others'. 'Others' examples would be less common head coverings, such as a chef hat, a football fan hat, a beret, among many others. Figure 4 presents images of different examples of these head coverings—the pictures in the two left columns are compliant religious options, while the pictures of the two rightmost columns show non-religious (thus non-compliant) options. We had also some skull caps, but we do not consider them on the list because they were not visible in the portraits.

We emphasize that, to consider a head covering as compliant, it is not enough that it is of a religious type. All head coverings must meet all the other requirements, such as being color-uniform and firm fitting. Figure 5 shows a pair of examples of religious head

Table 3 Different types of head coverings available and their number of occurrences in the DFIC-HC dataset

HC Type	Religious?	Total
No head covering	–	2770
Big Hat	×	324
Small Hat	×	252
Cap	×	723
Beanie	×	415
Ribbon	×	31
Glasses	×	42
Hijab	✓	813
White Gutra	✓	367
Brown Gutra	✓	104
Turban	✓	19
Others	✓ ×	899

× Non-religious head coverings. ✓ Religious head coverings



Fig. 3 A collection of the head coverings used in DFIC-HC dataset. To note that many participants had their own head coverings, which makes the total amount presented in the images even more diverse

coverings that would be classified as compliant (left), and two examples, on the right, of religious head coverings that, due to their ill-fitting, would be classified as not compliant.

3.4 Annotation strategy

The images in DFIC-HC were manually annotated by two independent people specially trained for this task. We provided an annotation tool where the task of the annotators was to select the compliance of each image regarding the head coverings requirements. The options were:

- 'Compliant', when all the requirements were met;



Fig. 4 An example of the available head coverings in the DFIC-HC dataset. The two left columns represent religious head coverings worn correctly while the two right ones comprise a collection of non-religious head coverings



Fig. 5 A collection of images representing how the placement of the religious head covering is essential for its compliance. The two left pictures are potentially compliant head coverings for woman and man that are fitted correctly, hence they might be considered compliant according to national rules. On the other hand, the two images on the right, although they are religious head coverings potentially compliant, their ill-fitment renders them as non-compliant

- 'No Way to Confirm', if for some reason (e.g., no light, occlusion) one could not verify the compliance;
- 'Non-compliant', where there were a specific options for each one of the main types of accessories considered (see the full list on Table 3);
- 'Other' when none of the other classes fits.

The conflicting labels were evaluated by a third expert annotator which decided the final label.

Each image label also includes information regarding the age and the country of origin of the corresponding subject. To obtain the results presented in this paper we grouped ages and origin countries according to the ones used by NIST, but after

the data are made available the research community can explore other more fine-grained analyses.

3.5 Ethical and legal aspects

3.5.1 Ethical aspects

In an increasingly global society where the automation of systems has been considered as crucial to increase their performance and efficiency, the automatic portrait analysis to check the compliance of portraits over the ICAO and ISO/IEC standards is not an exception. Such automatic validation processes can increase the velocity of MRTD issuance but can also allow for self-enrollment processes in some specific scenarios.

In this global society, different cultures, and specifically different religions, play a crucial role in the tension between people. Modern society rules are founded on the equality principles between people and persons and, consequently, different cultural and religious aspects, such like the use of head coverings, are of extremely importance to the inclusiveness of all, regardless of their origin and religious beliefs.

On the other hand, security of borders, or of any other controlled access, is also of a critical importance to society, to ensure that only authorized persons can pass a particular portal, of any kind. This way, persons may experience a sense of security and safety, not compatible with uncontrolled passages, opening the door to potential criminals or ill-intentioned persons. To be more concrete in terms of head coverings which is the object of this work, if one person wearing a head covering that partially covers their face, this could prevent automatic facial verification (1:1) or recognition (1:N) systems to match their live photo with the portrait stored in the MRTD document, if this stored portrait is not able to show the whole face without covers.

The two previous aspects can pose a significant challenge, between respect for religious liberty and security. As mentioned in the Introduction section, current automatic system for compliance check do not take into consideration religious aspects and limit their portrait analysis to the existence of hats or caps.

This paper then aims at improving inclusiveness of systems, by improving the religious-aware automatic ICAO and ISO/IEC standards of portraits regarding head coverings.

3.5.2 Legal aspects

We believe that authorized access to DFIC-HC for research purposes is crucial for the development of better solutions for the automatic ICAO requirements verification. Accordingly, we have followed all legal requirements that allow this access. All individuals in the photos have given explicit written permission for the use of their image rights and were shown and explained a privacy policy detailing how their images will be processed. Also, to ensure the effectiveness of people's rights, we have tagged each photo with a unique identifier assigned to each person. This pseudo anonymization technique prevents access to individuals' names by dataset users. The defined protocol ensures that only very few persons have access to biographical data, mainly for deleting or editing details, on demand of the individuals themselves. Finally, we contractually require all dataset users to comply with relevant data protection laws, including the GDPR. All the mentioned documents were carefully revised by a professional lawyer team.

4 Methods

The main objective of this section is to present a straightforward methodology to test the effectiveness of our novel dataset DFIC-HC. Although we found that our straightforward method to detect and classify compliance of portraits regarding religious-aware head coverings has outperformed the state-of-the-art methods (ICAOnet [13]) and OFIQ [14], we emphasize that our main focus was the dataset construction and its effectiveness assessment.

The constrained nature of the collected pictures allows the produced data to be well lit and of both high quality and resolution. In this case, and in opposition to an unconstrained scenario, a CNN-based model will be more skilled of producing discernible feature maps capable of distinguishing between face, background and head coverings. Henceforth, the aforementioned facts led us to a solution based on object detection principles within a deep learning framework.

In the field of object detection, the YOLO (You Only Look Once) network model [24] is currently highly regarded for its capability of detecting multiple objects in real-time. There are several available versions of the network, however, YOLOv8 is the one chosen. It is considered by the dev community to be the current state of the art by achieving higher mean Average Precision (mAP) on the COCO and Roboflow100 benchmarks [25] (not peer reviewed though). For instance, the second largest version of YOLOv8 can achieve the same mAP (52.9%) on the COCO benchmark as YOLOv7's largest model while needing much less parameters (43.7M vs. 71.3M) [26]. Therefore, we used YOLOv8 as our framework and it was further fine-tuned to achieve superior real-time performance and accuracy when deployed in facial image classification scenarios within ICAO's head coverings requirements.

With computational cost in mind without compromising accuracy, we opted for the second smallest variant of the classification network (6.4M parameters) and trained the model with a sub-set of 5932 images from the proposed DFIC-HC dataset, whereas 2973 are non-compliant and 2959 compliant regarding the head coverings requirement. For testing, 827 images balanced across categories were selected.

Our methodology follows the sequential pipeline proposed in Fig. 6. Firstly, all the data that comprise the dataset is processed. Using an implementation of MTCNN [27], the faces are detected and a loose crop (blue bounding box) that simulates real-world scenarios (i.e., a crop that would perform a token frontal portrait) is applied to both the train and test sets. Then, the images are resized to 256×256 and the model is fine-tuned as a binary classification model, over 10 epochs using AdamW [28] as the optimizer, with a learning rate of 0.01, weight decay 5×10^{-4} and batches of size 16. During inference, the model is deployed and tested over the above-stated 827 images. The classification scores computed by the trained network are thresholded to assert compliance, where a value of 0 reflects non-compliance and 1 compliance. This allows the extraction of the confusion matrix—True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN)—which in consequence facilitates the assessment of essential metrics such as the Accuracy, Half Total Error Rate (HTER) and the Equal Error Rate (EER). Accuracy can be defined as the overall percentage of correctly classified samples; HTER is respective to the average value between False Acceptance Rate (FAR) and False Rejection Rate (FRR), which in turn correspond to the rate at which non-compliant samples

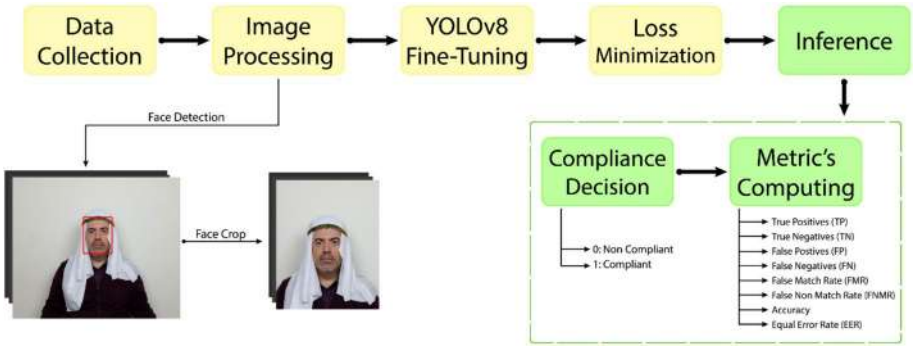


Fig. 6 A proposal of our methodology's sequential pipeline. All the images in the dataset are pre-processed before fine-tuning and further deploying of the model. This pre-process includes a facial detection stage that produces the bounding box (represented in red) that then serves as the base for a loose crop that produces the token frontal photo on the right. This crop's intention is to reproduce a real-world scenario for a MRTD's portrait

are classified as compliant, and to the rate at which compliant samples are classified as non-compliant, respectively. Finally, EER is a standard metric to evaluate the performance of biometric systems and can be defined as the rate at which both acceptance and rejections errors (FAR and FRR) are equal. The results presented in the following section, are respective to the acceptance threshold where the minimum value of HTER is reached, i.e., the best possible value for HTER, and the values of Accuracy, TP, TN, FP and FN are retrieved using this threshold.

4.1 Datasets used in experiments

For the following experiments, we used several datasets to conduct a variety of intra and inter-dataset testing protocols. Specifically, we have resorted to our proposed dataset as well as to annotated subsets of the AR [7] and FRGC [8]. The following subsections describe the used partitions for each dataset.

4.1.1 DFIC-HC

The training set for the created DFIC-HC dataset gathers 5932 images:

- 2959 compliant with all the head covering requirements;
- 2973 non-compliant with one or more head covering requirement.

All the methods are then tested on the testing set, which is also a sub-set of the complete DFIC-HC dataset, making sure that there was no repeated images from the training set. This set includes 827 images:

- 432 images compliant with all the head covering requirements;
- 395 images non-compliant with one or more head covering requirement.

Additionally, we also created a balanced version of this dataset for model training, **DFIC-HC (B)**. This training set is balanced in terms of age, gender, origin, as well as compliant and non-compliant, and it is composed by 2505 images (1269 compliant,

1236 non-compliant). The distribution across the 3 considered categories is presented in Fig. 7.

4.1.2 AR+FRGC

Current state-of-the-art method used images from datasets such as AR [7] and FRGC [8], as well as the FVC-onGoing benchmark carried out by Biolab [12]. To build a training and a test dataset that resemble the ones used by the authors of ICAONet [13], we gathered a sub-set of images from those datasets (respecting the same proportion). We manually annotated these images using the same methodology we did with DFIC-HC's images (see Sect. 3.4). Notice that there might be some differences in the annotations by other authors since, in our study, we distinguish when a head covering might be considered compliant because of religious circumstances, which was not considered before. In previous studies, including the state of the art ones, the considered head coverings were mainly hats and caps and any covering would be considered non-compliant.

The resulting train and test sets are then composed of:

- 3236 images for training—where 100 are non-compliant in terms of head coverings and 3136 are compliant;
- 143 images for testing—where 19 are non-compliant and 124 are compliant.

It is worth noticing that the resulting train and test datasets for this protocol are highly unbalanced, however, the full original datasets were indeed already unbalanced in terms of head coverings as previously pointed out too.

4.1.3 Mixed dataset

At last we wanted to evaluate the considered algorithms on a dataset containing a mix of images belonging from DFIC-HC and AR+FRGC. To do so we grouped all the images used for training and testing from the previous two datasets and ended up with:

- 9311 images for training—where 3154 are non-compliant in terms of head coverings and 6157 are compliant;
- 970 images for testing—where 414 are non-compliant and 556 are compliant.

4.2 Experiments

To evaluate the proposed approach to verify the compliance regarding head coverings, as well as the new data created, we defined several experiments to be conducted where we compared our results with the current state the art methods using many different training and testing sets, varying the demographic distribution of identities or the type of head covering used, so we can analyze inclusiveness. We also evaluate the impact of the pre-processing step of the proposed approach on the classification performance.

The first protocol (P1) aims at comparing our method with the state of the art alternative methods—'ICAONet' and 'OFIQ'—using data from the created DFIC-HC dataset for testing. Furthermore, we compare different versions of each method trained on DFIC-HC, 'AR+FRGC', and 'Mix' (joint group of DFIC-HC, AR and FRGC).

With the second protocol (P2) we evaluate our method and the alternatives when tested in other datasets—'AR+FRGC' and 'Mix'. P2.1 and P2.2 refer to the models trained in the different dataset domains and tested on AR+FRGC and Mix dataset, respectively.

The goal of the third protocol (P3) is to assess the performance of 'Ours' and 'ICAONet' methods when using a dataset balanced in terms of age, gender and origin during the training phase.

The following protocol (P4) aims at exploring how does the classification performance of both 'Ours' and 'ICAONet' methods (trained on 'DFIC-HC' or 'AR+FRGC') vary across different age groups, origins or genders.

Similarly, with the fifth protocol (P5), we explore how different kinds of head coverings (religious or not) influence the classification performance of the same models considered in P4.

Finally, in protocol 6 (P6), we want to evaluate the behavior of our approach without the image processing step, applied previously to the classification.

Table 4 summarizes all the conducted evaluation protocols.

5 Results

5.1 Protocol 1—general evaluation of compliance classification methods on DFIC-HC

We start by comparing our algorithm with the current state-of-the-art methods—ICAONet [13] and OFIQ [14] against DFIC-HC test set. In order to test intra and inter-dataset scenarios, a version of each method trained in DFIC-HC, 'AR+FRGC' and 'Mix' is evaluated on the created dataset, with the exception of OFIQ given that it is not possible to tune it for different domains. The DET curves and results are presented in Fig. 8 and Table 5, respectively.

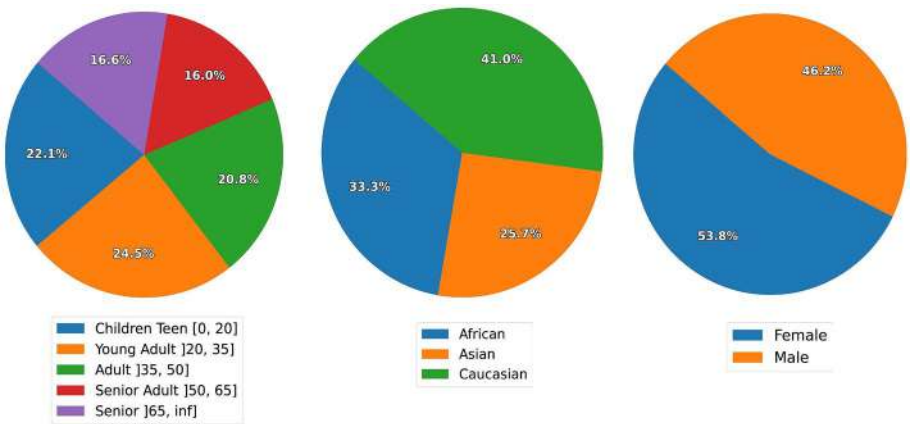


Fig. 7 Demographic distribution of a subset of DFIC-HC used for training that is balanced across the three considered variables: (left) age; (middle) ethnic origin; and (right) gender

Table 4 Description of the considered combinations of the experiments' protocols. Pn represents the n-th protocol

	Method	Trained with	Protocol						
			P1	P2		P3	P4	P5	P6
				2.1	2.2				
Model	Ours	DFIC-HC	✓	✓	✓	–	✓	✓	–
		DFIC-HC (B)	–	–	–	✓	–	–	–
		AR+FRGC	✓	✓	✓	–	✓	✓	–
		Mix	✓	✓	✓	–	–	–	–
	Ours (no pre-processing)	DFIC-HC	–	–	–	–	–	–	✓
	ICAONet	DFIC-HC	✓	✓	✓	–	✓	✓	–
		DFIC-HC (B)	–	–	–	✓	–	–	–
		AR+FRGC	✓	✓	✓	–	✓	✓	–
		Mix	✓	✓	✓	–	–	–	–
	OFIQ	–	✓	✓	✓		–	–	–
Tested on	DFIC-HC		✓	–	–	✓	–	–	✓
	AR+FRGC		–	✓	–	–	–	–	–
	Mix		–	–	✓	–	–	–	–
	Demographic subsets		–	–	–	–	✓	–	–
	Head coverings subsets		–	–	–	–	–	✓	–

The "Model" rows indicate the algorithm chosen and where the model was trained. Furthermore, DFIC-HC (B) refers to a balanced subset of the original DFIC-HC and "Mix" is the joint group of DFIC-HC, AR and FRGC

Table 5 Protocol 1—performance of the different models trained on different datasets and tested on the DFIC-HC subset

Method	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**OFIQ**
Train	**DFIC-HC**		**AR+FRGC**		**Mix**		**-**
Test	**DFIC-HC**						
EER	**10.59%**	26.64%	27.27%	49.54%	10.81%	36.18%	19.58%
HTER	**8.53%**	25.94%	26.70%	45.67%	10.04%	33.30%	15.95%
Accuracy	**91.05%**	74.37%	72.91%	56.83%	89.72%	67.71%	83.43%
#TP	381	347	302	392	385	357	304
#TN	372	268	301	78	357	203	386
#FP	11	115	82	305	26	180	9
#FN	63	97	142	52	59	87	128
t	0.7	0.97	0.88	0.07	0.68	0.97	0.96

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

"Mix" is a dataset constituted by the DFIC-HC, AR and FRGC sets. HTER is the half total error rate; TP, TN, FP, FN relate, respectively, to the number of true positives, true negatives, false positives and false negatives; and *t* is the acceptance threshold

5.1.1 Analysis of methods trained on DFIC-HC (Intra-Dataset)

For a better comparison between our approach and remaining methods, their performance was evaluated on a validation set along a range of acceptance thresholds and the metrics are reported where they individually perform better in HTER.

Table 5 summarizes the results obtained in terms of EER, HTER, accuracy, number of true positives, true negatives, false positives and false negatives. By positive, we refer to compliant samples and by negative we refer to non-compliant samples.

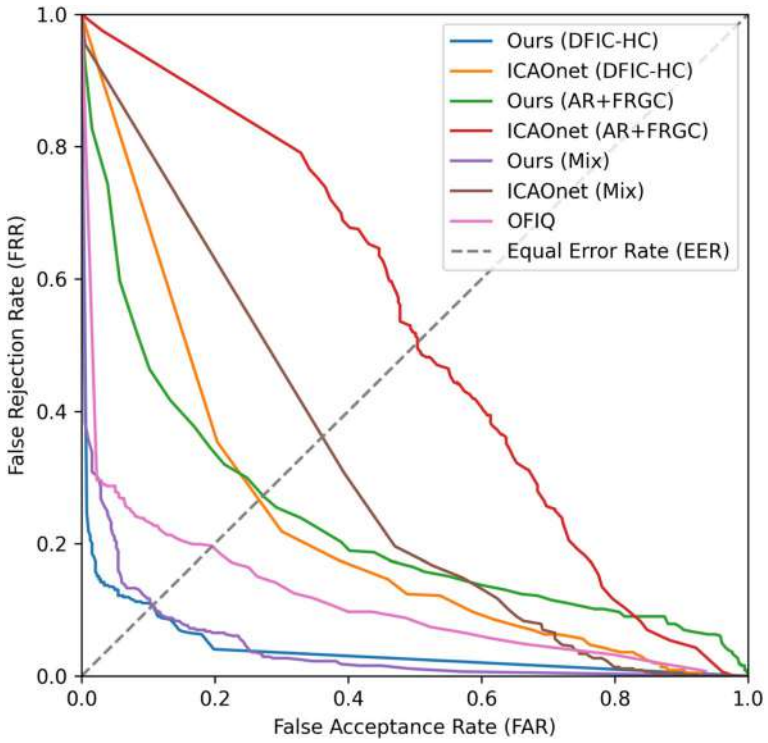


Fig. 8 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets. All the combinations were tested on the same DFIC-HC test set (**Protocol 1**)

Looking at the first two columns on the left, we can confirm that overall our method has higher frequency than ICAONet for the true predictions (TP and TN) and the lowest for false predictions (FP and FN), which is obviously reflected in the accuracy and HTER results where our model achieves higher performance. Looking at the HTER specifically our method significantly outperforms the alternative, with an HTER_Ours = 8.53% compared to an HTER_ICAONet = 25.94%, which can also be confirmed by comparing the blue and orange curves in Fig. 8. The improvement is particularly significant in the case of non-compliant samples (images with different types of non-compliant head coverings), which our method detects more frequently (TN = 372; FN = 63) than ICAONet (TN = 268; FN = 97).

On the other hand, when comparing our method to OFIQ (last column), the latter shows slightly worse results with an HTER_OFIQ = 15.95% on the complete DFIC-HC test set, but still a good performance considering the unseen domain of our dataset, which confirms its great relevance in this context. However, when looking in particular to the images with compliant head coverings, OFIQ fails in most of the images, which in turn leads to 128 False Negatives. This is because it does not consider any head covering as compliant unlike our approach that gets half the number of FN, by indeed considering compliant most of the religious cases.

5.1.2 Analysis of methods trained on AR+FRGC (Inter-Dataset)

Table 5 shows the results obtained after training both our method and ICAOnet in this new dataset (AR+FRGC) but still testing on images from DFIC-HC—see columns 3 and

4. Before extracting the metrics, we analyzed which acceptance threshold would lead to better results for each method as before. The results have shown that our method significantly outperforms ICAONet again, with an HTER_Ours = 26.70% versus HTER_ICAONet = 45.67%. In Fig. 8, one can confirm the performance difference between the two methods by inspecting the green and red curves. Additionally, the results are much worse when tested on images from DFIC-HC, which we expected since the images used during testing were much different from the ones used during training.

5.1.3 Analysis of methods trained on mixed dataset

The columns 5 and 6 on Table 5 show the results obtained after training both our method and ICAOnet on this new mixed dataset but still testing on images from DFIC-HC. We performed the acceptance threshold analysis once again to compare the best possible individual results. Our method shows significantly better results, repeating the previous performance, with an HTER_Ours = 10.04% versus HTER_ICAONet = 33.30%. Again, this result is worse than the one obtained when training and testing with images from DFIC-HC, which is expected, but the difference is subtle (see blue and purple curves in Fig. 8).

5.2 Protocol 2—evaluation on other datasets

Having confirmed that our pipeline performs better in the images from the proposed DFIC-HC dataset, we extended the evaluation to other datasets to confirm if the results remain consistent.

5.2.1 Testing on AR+FRGC

As for the test on the AR+FRGC dataset, we tested the seven previous models—Ours trained on DFIC-HC; Ours trained on AR+FRGC; Ours trained on Mix; ICAONet trained on DFIC-HC; ICAONet trained on AR+FRGC; ICAONet trained on Mix; and OFIQ—but this time on a new test set with images belonging to AR+FRGC. We present all the results in Table 6, and the detection error trade-off curves illustrated in Fig. 9. Our method proved to significantly outperform the ICAONet alternatives in this new test set, regardless of the training images. As expected, the best results are those corresponding to the methods trained in the Mix dataset, since it allows the methods to explore the diversity present in DFIC-HC and adapt to the specific domain of AR+FRGC. Regarding the remaining columns, the results are better when the models are trained on the same type of images as the ones used for testing (AR+FRGC), as expected. However, the difference is subtle when comparing the intra- and inter-dataset scenarios (particularly regarding our method with HTER = 16.35% vs. 12.95%), which confirms a better generalization ability of DFIC-HC dataset.

Regarding the comparison of Ours method with OFIQ, the latter presents slightly better results (HTER_OFIQ = 5.26% vs. HTER_Ours_Mix = 9.08%). This may be due to the fact that AR and FRGC are well-established datasets in this field, and it is highly probable that OFIQ is already highly tuned to the domains of these particular datasets, hence giving better results.

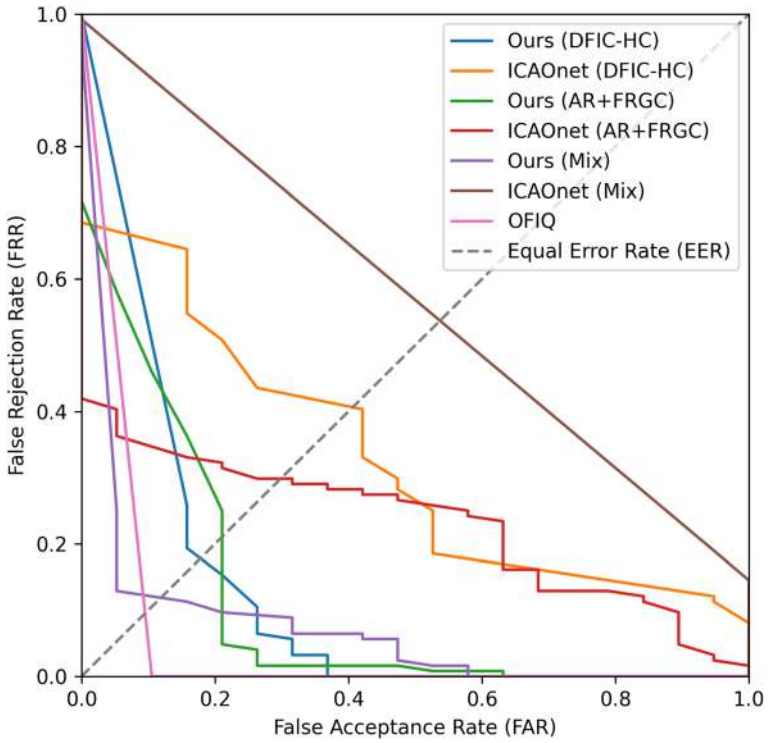


Fig. 9 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets. All the combinations were tested on the same AR+FRGC test set (**Protocol 2.1**)

Table 6 Protocol 2.1—performance of the different models trained on different datasets and tested on the test partition of AR and FRGC datasets

Method	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**OFIQ**
Train	**DFIC-HC**		**AR+FRGC**		**Mix**		**–**
Test	**AR+FRGC**						
EER	17.81%	40.63%	25.11%	29.84%	11.89%	53.72%	**9.75%**
HTER	16.35%	34.93%	12.95%	20.78%	9.08%	49.60%	**5.26%**
Accuracy	90.01%	58.74%	93.01%	67.83%	88.11%	13.99%	**98.60%**
#TP	116	70	118	79	108	1	124
#TN	14	14	15	18	18	19	17
#FP	5	5	4	1	1	0	2
#FN	8	54	6	45	16	123	0
t	0.93	0.89	0.86	0.93	0.95	0.99	0.46

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

"Mix" is a dataset constituted by the DFIC-HC, AR and FRGC sets. HTER is the half total error rate; TP, TN, FP, FN relate, respectively, to the number of true positives, true negatives, false positives and false negatives; and *t* is the acceptance threshold

5.2.2 Testing on a mixed dataset

To conclude the evaluation of the methods regarding the datasets used for training and testing, we repeated all the previous models—Ours and ICAONet trained on the three different datasets (DFIC-HC, AR+FRCG and the mixed alternative) and OFIQ—but this time on a test set including images from both DFIC-HC and

Table 7 Protocol 2.2—performance of the different models trained on different datasets and tested on the test partition of the "Mix" dataset

Method	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**Ours**	**ICAONet**	**OFIQ**
Train	**DFIC-HC**		**AR+FRGC**		**Mix**		**-**
Test	**Mix**						
EER	**9.26%**	32.94%	24.67%	47.71%	10.60%	36.53%	16.18%
HTER	**7.86%**	32.09%	24.16%	45.06%	9.79%	33.29%	12.84%
Accuracy	**91.65%**	67.32%	75.36%	60.72%	89.90%	69.48%	85.67%
#TP	507	366	415	504	502	471	428
#TN	382	287	316	85	370	203	403
#FP	20	115	86	317	32	199	11
#FN	61	202	153	64	66	97	128
t	0.60	0.97	0.88	0.08	0.70	0.97	0.96

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

"Mix" is a dataset constituted by the DFIC-HC, AR and FRGC sets. HTER is the Half total error rate; TP, TN, FP, FN relate, respectively, to the number of true positives, true negatives, false positives and false negatives; and *t* is the acceptance threshold

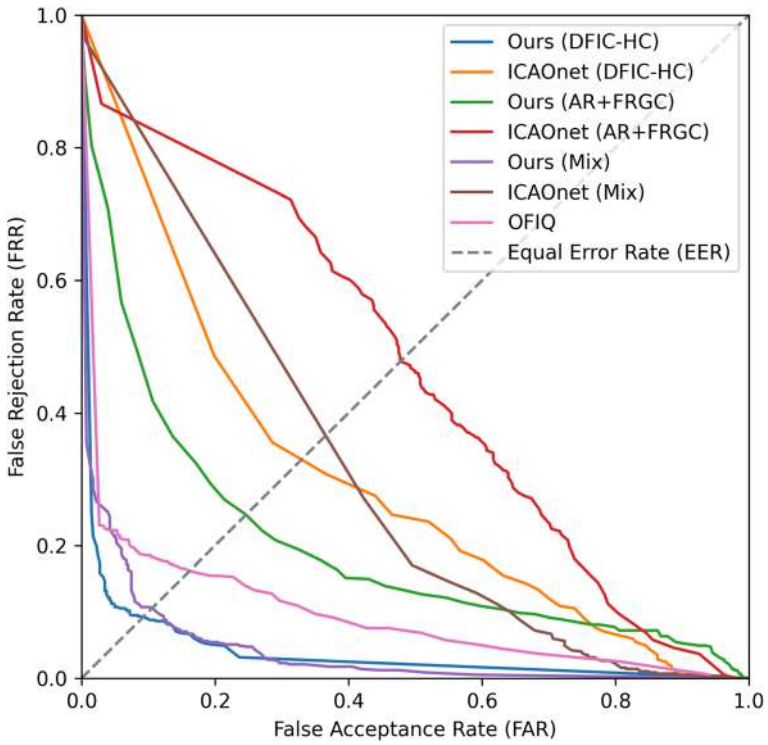


Fig. 10 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets. All the combinations were tested on the same Mixed test set (**Protocol 2.2**)

AR-FRGC, which is the most diverse test set—see the results in Table 7 and DET curves in Fig. 10.

Our method significantly outperformed ICAONet on all tests, which now allows us to say that it is the best option, considering these two options, to verify the compliance with the ICAO requirements regarding head coverings, regardless of the dataset.

Regarding the OFIQ method, once again its performance was hindered by the presence of compliant head coverings, maintaining the number of FN=128 as in the first protocol.

However, an interesting result is that in this case the best result was not achieved when training and testing with images from the same dataset, but rather when training on DFIC-HC and testing on the mixed dataset (HTER = 7.86%, leftmost column of Table 7). This happens possibly because the mixed test set has a lower percentage of images with head coverings (and less varied), which makes it an easier test set than the one used on Protocol 1 with images only from DFIC-HC. This suggests that besides having better generalization ability, the results obtained with our new DFIC-HC dataset are also more robust.

5.3 Protocol 3—balanced versus non-balanced data

In the previous Results sections the training sets were unbalanced in terms of age, gender and origin because we wanted to maximize the number of images with compliant and non-compliant head coverings to make the model learn the best it could. However, for it to generalize in an unbiased way, we defined a training balanced in terms of age, gender, origin, as well as compliant and non-compliant images.

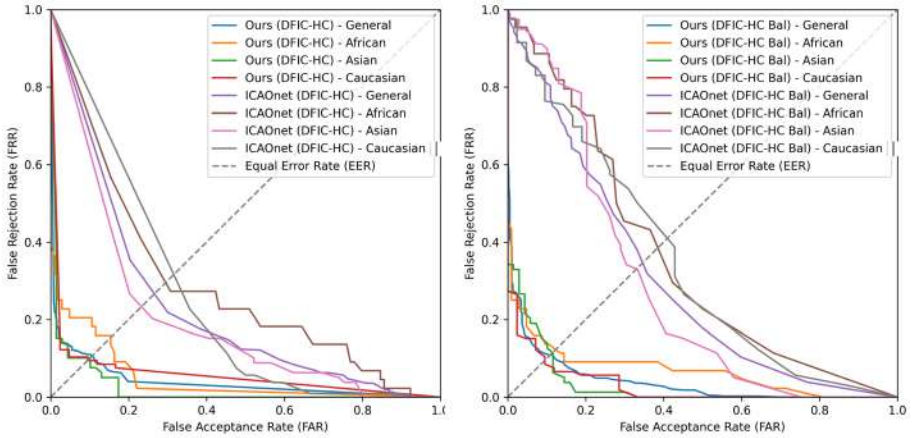
We tested the new versions of both our model and ICAONet with the balanced training on the same DFIC-HC test set used in the previous protocols (making sure to filter the images that were now included on the new balanced training set). The results are summarized on Table 8. There is a slight increase in the HTER when comparing to the one obtained with the unbalanced train previously reported on Table 5 (HTER_Ours_Balanced = 10.00% vs. HTER_Ours_Unbalanced = 8.53%), which we expected regarding that the number of images used for training also decreased to maintain the balanced across the 3 considered categories. However, this result is still significantly better than the one obtained with ICAONet (HTER_ICAONet_Balanced = 33.71%) and considerably low so we can say that it performs well and in an inclusive and unbiased way. Table 8

Table 8 Protocol 3—results of both the algorithms trained on a balanced subset of the DFIC-HC and tested on a filtered version of the test partition of DFIC-HC and its subsets regarding ethnic origin

Method	**Ours**				**ICAONet**			
Train	**DFIC-HC (balanced)**							
Test	**DFIC-HC**	**African**	**Asian**	**Caucasian**	**DFIC-HC**	**African**	**Asian**	**Caucasian**
EER	10.88%	12.28%	11.47%	**10.04%**	34.65%	38.55%	33.06%	41.01%
HTER	10.00%	11.32%	**8.96%**	9.01%	33.71%	35.93%	28.52%	35.83%
Accuracy	89.99%	90.54%	**91.22%**	**91.22%**	66.32%	61.49%	72.30%	68.24%
# TP	333	37	74	97	289	31	66	78
# TN	359	97	61	38	221	60	41	23
# FP	24	7	8	4	162	44	28	19
# FN	53	7	5	9	97	13	13	28
t	0.38	0.33	0.27	0.35	0.03	0.02	0.04	0.04

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

This filtering was achieved by removing repeated samples between training and testing. HTER is the half total error rate; TP, TN, FP, FN relate, respectively, to the number of true positives, true negatives, false positives and false negatives; and *t* is the acceptance threshold



(a) Unbalanced DFIC-HC Training set. (b) Balanced DFIC-HC Training set.

Fig. 11 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets, specifically: **a** Ours and ICAONet trained on the original DFIC-HC; and **b** Ours and ICAONet trained on a balanced version of DFIC-HC. All the combinations were tested on the same DFIC-HC sub-test sets (**Protocol 3**)

Table 9 Protocol 4 (Age)—results of the models trained on all the datasets and tested on subsets of the test partition of DFIC-HC aimed at testing the sensibility of the model over different age groups

	Ours (DFIC-HC)		**ICAONet (DFIC-HC)**		**Ours (AR+FRGC)**		**ICAONet (AR+FRGC)**	
Children/teen	**3.91%**	*3.91%*	25.30%	*19.57%*	25.99%	*44.04%*	36.02%	*33.89%*
Young adult	**8.04%**	*4.46%*	40.46%	*31.72%*	31.02%	*29.32%*	54.46%	*48.21%*
Adult	**7.10%**	*7.10%*	21.77%	*21.52%*	34.51%	*32.75%*	47.35%	*45.12%*
Senior adult	**9.00%**	*9.00%*	38.00%	*34.00%*	32.00%	*29.00%*	48.00%	*46.00%*
Senior	**16.42%**	*16.42%*	41.27%	*37.71%*	24.40%	*20.31%*	48.65%	*44.04%*
Baseline HTER	8.53%		25.94%		26.70%		45.67%	

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The italic values correspond to the best possible value for each class in terms of HTER while the remaining values are extracted with the threshold previously defined in Table 5. The baseline values represent the general HTER when the models were evaluated in the global DFIC-HC test set, i.e., not divided in terms of age

also shows the results obtained across the three origin groups considered—African, Asian and Caucasian. The differences between these groups are very subtle, which again confirms that the balanced training leads to unbiased results, which can be confirmed in Fig. 11, where the poorer performance in the African group using the original DFIC-HC train set (orange curve in Fig. 11a) can be mitigated using a balanced training set (orange curve in Fig. 11b).

5.4 Protocol 4—demographic effects

With this protocol we want to analyze the impact of the age, gender and origin in the obtained results. We thus computed the performance metric across all classes of each of the three variables being analyzed. In Appendix A is presented one detection error trade-off (DET) figure for each variable.

Table 10 Protocol 4 (Ethnicity)—results of the models trained on all the datasets and tested on subsets of the test partition of DFIC-HC aimed at testing the sensibility of the model over different ethnic origins

	Ours (DFIC-HC)		ICAONet (DFIC-HC)		Ours (AR+FRGC)		ICAONet (AR+FRGC)	
African	**12.81%**	*12.63%*	31.99%	*29.02%*	47.47%	*42.00%*	49.48%	*47.55%*
Asian	**8.32%**	*7.24%*	23.44%	*23.17%*	38.54%	*32.32%*	51.60%	*48.07%*
Caucasian	**7.57%**	*7.32%*	28.28%	*27.58%*	22.78%	*21.61%*	43.33%	*43.33%*
Baseline HTER	8.53%		25.94%		26.70%		45.67%	

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The italic values correspond to the best possible result for each class in terms of HTER while the remaining values are extracted with the threshold previously defined in Table 5. The baseline values represent the general HTER when the models were evaluated in the global DFIC-HC test set, i.e., not divided in terms of ethnic origin

Table 11 Protocol 4 (Gender)—results of the models trained on all the datasets and tested on a subset of the test partition of DFIC-HC aimed at testing the sensibility of the model over the considered genders

	Ours (DFIC-HC)		ICAONet (DFIC-HC)		Ours (AR+FRGC)		ICAONet (AR+FRGC)	
Female	**3.65%**	*3.13%*	17.60%	*17.60%*	30.43%	*27.74%*	37.66%	*37.66%*
Male	**12.66%**	*12.66%*	29.95%	*29.95%*	29.10%	*27.86%*	47.98%	*47.10%*
Baseline HTER	8.53%		25.94%		26.70%		45.67%	

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The italic values correspond to the best possible value for each class in terms of HTER while the remaining values are extracted with the threshold previously defined in Table 5. The baseline values represent the general HTER when the models were evaluated in the global DFIC-HC test set, i.e., not divided in terms of gender

Starting with the age (see Table 9) it can be observed that the best results are obtained with the 'Children/Teen' group ([0,20]) - HTER = 3.91%. This might be explained due to this group being the second with the largest amount of images in DFIC-HC dataset, combined with being the group whose participants vary less in terms of age, being more concentrated towards the higher end of the age interval, which decreases the variety of the images. Participants from the first half of the age interval (0 to 10 years old) were much more difficult to obtain. On the other hand, the worst results are obtained with the 'Senior' group ([> 65]), because, inversely to the 'Children/Teen' group, this was the age group with fewer participants and therefore with less samples for the model to learn. But overall our method outperforms ICAONet regardless of the age group considered or the dataset used for training.

Regarding the origin of the participants (see Table 10), the Caucasian group leads to the best results (HTER_Caucasian = 7.32%) and the less accurate results are obtained for the African group (HTER_African = 12.63%). This effect seems to be caused by lower contrast in some images with African participants, due to the darker skin color leading to worse automatic white balancing by the camera. The contrast with the head accessories could also be lower in many cases. However, regardless of the group or the training set considered, our method presents again the lowest HTER value.

Looking at the gender effect on the results presented in Table 11, one can notice a clear increase on the HTER when considering male participants (HTER_Male = 12.66%)

Table 12 Protocol 5—a collection of HTER values for all the algorithms and training combinations computed from a division of the test partition of DFIC-HC dataset in terms of the different head coverings available

	Religious?	Ours (DFIC-HC)		ICAONet (DFIC-HC)		Ours (AR+FRGC)		ICAONet (AR+FRGC)	
No HC	–	**1.34%**	*0.33%*	6.02%	*0.00%*	4.18%	*0.00%*	2.34%	*0.00%*
Hats	×	**0.00%**	*0.00%*	5.80%	*0.00%*	7.97%	*0.00%*	44.20%	*2.17%*
Cap	×	**0.70%**	*0.00%*	16.43%	*0.00%*	4.55%	*0.00%*	39.86%	*1.05%*
Beanie	×	**2.27%**	*0.00%*	18.18%	*0.00%*	6.83%	*0.00%*	34.09%	2.27%
Ribbon	×	25.00%	*0.00%*	**0.00%**	*0.00%*	25.00%	*0.00%*	25.00%	*0.00%*
Glasses	×	**0.00%**	*0.00%*	**0.00%**	*0.00%*	0.00%	*0.00%*	50.00%	*0.00%*
Hijab	✓	**3.49%**	*1.16%*	13.95%	*0.00%*	31.40%	*0.00%*	16.28%	*0.00%*
Gutra	✓	**44.36%**	41.96%	53.53%	*46.45%*	53.02%	*45.39%*	52.16%	*46.39%*
Turban	✓	5.88%	*0.00%*	**2.94%**	*0.00%*	26.47%	*0.00%*	29.41%	*0.00%*
Other	✓/×	**4.07%**	*3.45%*	43.67%	*39.74%*	71.96%	*50.00%*	47.48%	*44.42%*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The "Hats" class comprises both of the "Big Hat" and "Small Hat" categories and, similarly, the "Gutra" class joins the "White Gutra" and "Brown Gutra" as seen in Table 3. The values in italic indicate the best possible HTER for that class. × Non-religious head coverings. ✓ Religious head coverings

when compared with female participants (HTER_Female = 3.13%), for the compliance on head coverings. This can be explained by male participants having more facial hair, what in many cases can be confused by the model as some type of non-compliant covering. Also, images from bald participants (that in the examined datasets are only male) could also lead to classification inaccuracies.

5.5 Protocol 5—type of head covering effect

In this protocol, we explore the effect of the various types of head coverings we have in our dataset on the model performance and how they compare to ICAONet—Table 12. In Appendix A there is one table for each head covering type with more detailed metrics, in addition with the DET curves for the different configurations in Fig. 15.

When looking at the non-religious types of head coverings, we can consider that both methods already have an almost perfect classification ability with very low HTERs, regardless of the dataset used for training.

However, when looking at the religious coverings there are much more clear differences. The most difficult head coverings to classify are the gutras (Arabic religious male coverings), with an HTER_Gutra for our images equal to 41.96%. And if we look at Table 20, we can see that the number of false positives doubles the number of false negatives, which means the model mainly fails at detecting non-compliant gutras. Basically, it classifies almost all gutras as compliant. Nonetheless, when comparing with the ICAONet results (trained on DFIC-HC as well as on the AR+FRGC) our method performs better. Overall our method outperforms the ICAONet method at all types of head coverings, but the improvement is particularly clear in the religious coverings, especially in the 'Other' group (although it includes non-religious coverings as well), where there is a difference in the HTER obtained when training with our images of more than 35pp (HTER_Ours_DFIC = 3.45% vs. HTER_ICAONet_DFIC = 39.74%). When trained on AR+FRGC there is a very obvious decrease in performance (HTER_Ours_AR+FRGC =

50.00%, which is expected knowing the AR and the FRGC datasets lack images with religious coverings and therefore do not allow the model to learn how to classify this type of accessories). This confirmation in the obtained results is particularly relevant because it was one of the main motivations for the creation of DFIC-HC dataset.

5.6 Protocol 6—image pre-processing effect

To finish the Results section, we conducted a final ablation study regarding the proposed algorithm directly. We tested the effect of the 'Image Processing' step illustrated in Fig. 6, in which we run a face detector and we crop the image.

To do so we run our algorithm without the referred step on the testing set of 827 images defined in Protocol 1 and obtained an HTER = 15.05%, which almost doubles the HTER result obtained including the pre-processing step (8.53%), thus confirming its relevance.

6 Conclusion

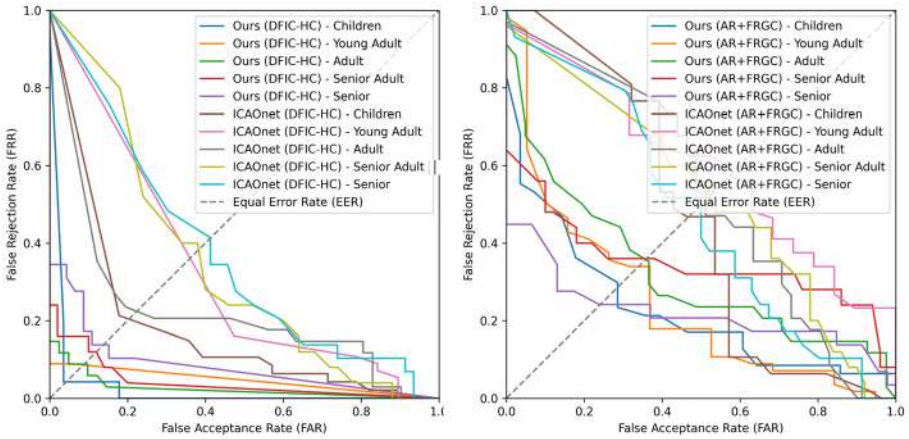
This work makes significant contributions in the form of a dataset and pipeline aimed at automating the verification of compliance with ICAO requirements concerning the presence of head coverings in facial images used in machine-readable travel documents. The existing methods fail to consider the acceptance of certain head coverings based on religious or cultural reasons, focusing primarily on the detection of hats or caps. In contrast, our approach specifically extends these considerations to include cases of religious coverings that can be considered compliant. To support our approach, we created a new dataset consisting of 6759 facial images representing 910 diverse and balanced identities and accommodating the inclusion of more accessory types such as compliant and non-compliant religious coverings, which was lacking in all the datasets available in the literature. Using this dataset, we fine-tuned and trained a classification model based on the YOLOv8 framework, resulting in a performance that significantly outperforms the current state of the art results when religious coverings are present. We confirmed it through a series of protocols that evaluated our model in many different scenarios, against the state-of-the-art methods—ICAONet and OFIQ—where we vary the dataset used during training and testing, the balance of the dataset and also tests where we go across demographic group (according to age, gender and origin) and the type of head covering used.

This work highlights the lack of inclusiveness of religious factors when verifying compliance with head covering requirements, and demonstrates the efficacy of our approach in accurately identifying compliant religious-aware head coverings.

In the future, we plan to submit our model to the FVC-onGoing benchmark carried out by Biolab [12] to be able to test it with one more dataset particularly relevant in this context of ICAO requirements. Furthermore, we would like to extend our dataset to include examples of head covering that cannot be removed because of medical reasons, which along with the religious coverings is stated by ICAO and ISO/IEC standards as able to be considered compliant.

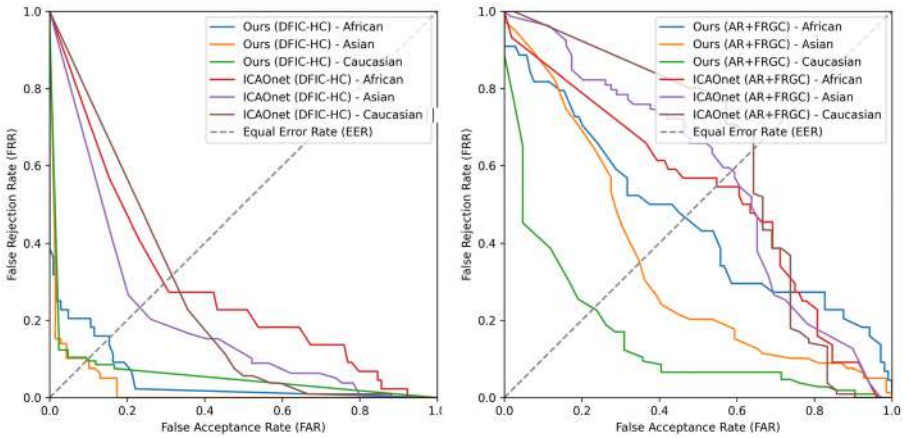
A Appendix

See Figs. 12, 13, 14 and 15, Tables 13, 14, 15, 16, 17, 18, 19 and 20, 21, 22.



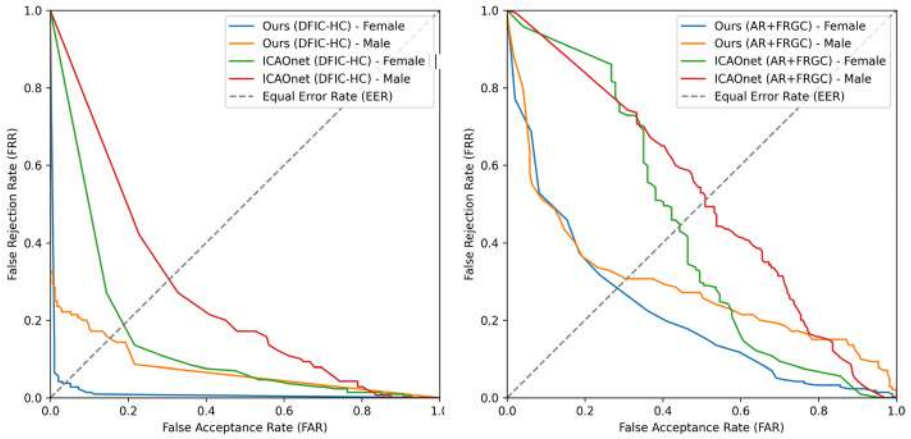
(a) DFIC-HC Training set. (b) AR+FRGC Training set.

Fig. 12 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets, specifically: (**a**) Ours and ICAONet trained on DFIC-HC; and (**b**) Ours and ICAONet trained on AR+FRGC. All the combinations were tested on the same DFIC-HC age sub-test sets (**Protocol 4 (Age)**)



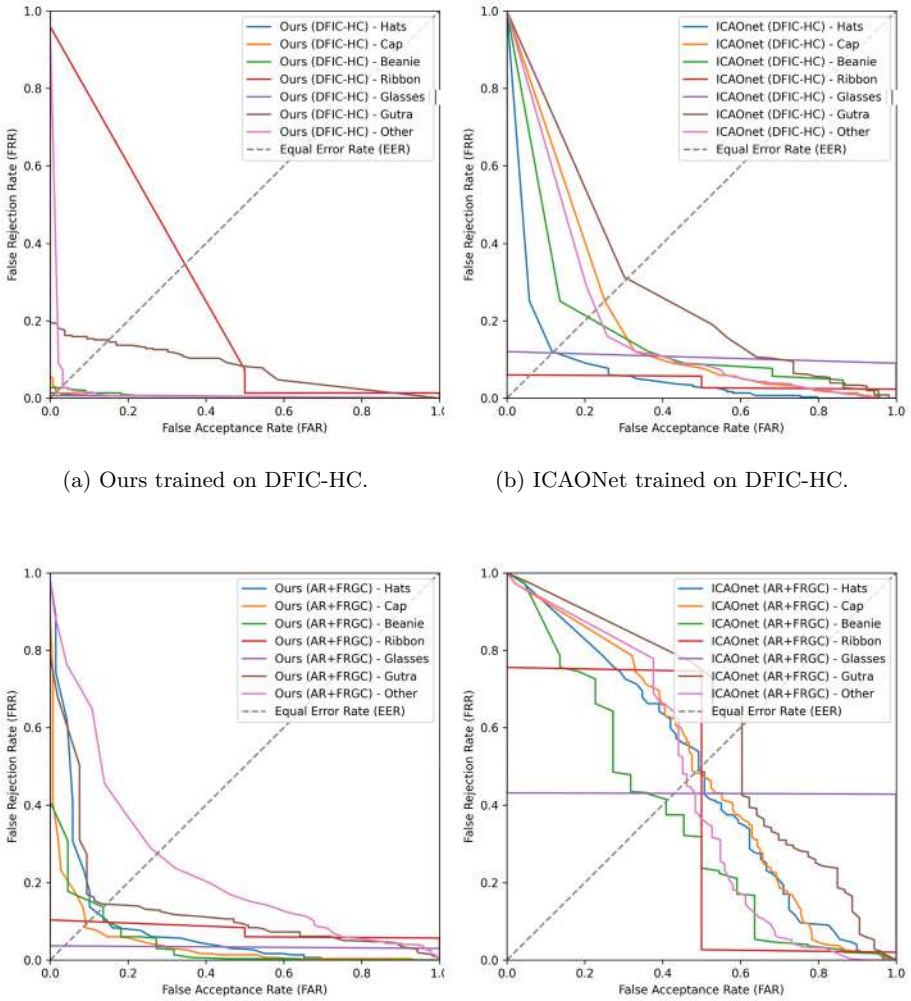
(a) DFIC-HC Training set. (b) AR+FRGC Training set.

Fig. 13 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets, specifically: (**a**) Ours and ICAONet trained on DFIC-HC; and (**b**) Ours and ICAONet trained on AR+FRGC. All the combinations were tested on the same DFIC-HC ethnicity sub-test sets (**Protocol 4 (Ethnicity)**)



(a) DFIC-HC Training set. (b) AR+FRGC Training set.

Fig. 14 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets, specifically: (**a**) Ours and ICAONet trained on DFIC-HC; and (**b**) Ours and ICAONet trained on AR+FRGC. All the combinations were tested on the same DFIC-HC gender sub-test sets (**Protocol 4 (Gender)**)



(a) Ours trained on DFIC-HC.

(b) ICAONet trained on DFIC-HC.

(c) Ours trained on AR+FRGC.

(d) ICAONet trained on AR+FRGC.

Fig. 15 Detection error tradeoff (DET) curves computed across a range of acceptance thresholds for different models and training sets. To compute the DET curves, only the head coverings with non-compliant samples, with addition with the compliant samples, were used (**Protocol 5**)

Table 13 Protocol 5 (No HC)—detailed results regarding the "No HC" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (No HC)**							
HTER	**1.34%**	*0.33%*	6.02%	*0.00%*	4.18%	*0.00%*	2.34%	*0.00%*
Accuracy	**97.32%**	*99.33%*	87.96%	*100.0%*	91.64%	*100.0%*	95.32%	*100.0%*
# True positives	291	*297*	263	*299*	274	*299*	285	*299*
# True negatives	0	*0*	0	*0*	0	*0*	0	*0*
# False positives	0	*0*	0	*0*	0	*0*	0	*0*
# False negatives	8	*2*	36	*0*	25	*0*	14	*0*
Acceptance threshold	0.70	*0.01*	0.97	*0.52*	0.88	*0.59*	0.07	*0.01*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 14 Protocol 5 (Hats)—detailed results regarding the "Hats" category (comprising the "Big Hat" and "Small Hat" classes) from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Hats)**							
HTER	**0.00%**	*0.00%*	5.80%	*0.00%*	7.97%	*0.00%*	44.20%	*2.17%*
Accuracy	**100.0%**	*100.0%*	88.41%	*100.0%*	84.06%	*100.0%*	11.59%	*95.65%*
# True positives	0	*0*	0	*0*	0	*0*	0	*0*
# True negatives	69	*69*	61	*69*	58	*69*	8	*66*
# False positives	0	*0*	8	*0*	11	*0*	61	*3*
# False negatives	0	0	0	*0*	0	*0*	0	*0*
Acceptance threshold	0.70	*0.99*	0.97	*0.99*	0.88	*0.99*	0.07	*0.99*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 15 Protocol 5 (Cap)—detailed results regarding the "Cap" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Cap)**							
HTER	**0.70%**	*0.00%*	16.43%	0.00%	4.55%	0.00%	39.86%	1.05%
Accuracy	**98.60%**	*100.0%*	67.13%	*100.0%*	90.91%	*100.0%*	20.28%	*97.90%*
# True positives	0	*0*	0	*0*	0	*0*	0	*0*
# True negatives	141	*143*	96	*143*	130	*143*	29	*140*
# False positives	2	*0*	47	*0*	13	*0*	114	*3*
# False negatives	0	*0*	0	*0*	0	*0*	0	*0*
Acceptance threshold	0.70	*0.99*	0.97	*0.99*	0.88	*0.99*	0.07	*0.99*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 16 Protocol 5 (Beanie)—detailed results regarding the "Beanie" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Beanie)**							
HTER	**2.27%**	*0.00%*	18.18%	0.00%	6.83%	*0.00%*	34.09%	*2.27%*
Accuracy	**95.45%**	*100.0%*	63.64%	*100.0%*	86.36%	*100.0%*	31.82%	*95.45%*
# True positives	0	*0*	0	*0*	0	*0*	0	*0*
# True negatives	21	*22*	14	*22*	19	*22*	7	*21*
# False positives	1	*0*	8	*0*	3	*0*	15	*1*
# False negatives	0	*0*	0	*0*	0	*0*	0	*0*
Acceptance threshold	0.70	*0.99*	0.97	*0.99*	0.88	*0.99*	0.07	*0.99*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 17 Protocol 5 (Ribbon)—detailed results regarding the "Ribbon" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Ribbon)**							
HTER	25.00%	*0.00%*	**0.00%**	*0.00%*	25.00%	*0.00%*	25.00%	*0.00%*
Accuracy	50.00%	*100.0%*	**100.0%**	*100.0%*	50.00%	*100.0%*	50.00%	*100.0%*
# True positives	0	*0*	0	*0*	0	*0*	0	*0*
# True negatives	1	*2*	2	*2*	1	*2*	1	*2*
# False positives	1	*0*	0	*0*	1	*0*	1	*0*
# False negatives	0	*0*	0	*0*	0	*0*	0	*0*
Acceptance threshold	0.70	*0.99*	0.97	*0.99*	0.88	*0.99*	0.07	*0.99*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 18 Protocol 5 (Glasses)—detailed results regarding the "Glasses" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Glasses)**							
HTER	0.00%	*0.00%*	0.00%	*0.00%*	0.00%	*0.00%*	50.00%	*0.00%*
Accuracy	100.0%	*100.0%*	100.0%	*100.0%*	100.0%	*100.0%*	0.00%	*100.0%*
# True positives	0	*0*	0	*0*	0	*0*	0	*0*
# True negatives	1	*1*	1	*1*	1	*1*	0	*1*
# False positives	0	*0*	0	*0*	0	*0*	1	*0*
# False negatives	0	*0*	0	*0*	0	*0*	0	*0*
Acceptance threshold	0.70	*0.99*	0.97	*0.99*	0.88	*0.99*	0.07	*0.99*

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 19 Protocol 5 (Hijab)—detailed results regarding the "Hijab" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Hijab)**							
HTER	**3.49%**	*1.16%*	13.95%	*0.00%*	31.40%	*0.00%*	16.28%	0.00%
Accuracy	**93.02%**	*97.67%*	72.09%	*100.0%*	37.21%	*100.0%*	67.44%	*100.0%*
# True positives	40	*42*	31	*43*	16	*43*	29	*43*
# True negatives	0	*0*	0	*0*	0	*0*	0	*0*
# False positives	0	*0*	0	*0*	0	*0*	0	*0*
# False negatives	3	*1*	12	*0*	27	*0*	14	*0*
Acceptance threshold	0.70	*0.11*	0.97	*0.77*	0.88	*0.50*	0.07	*0.00*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 20 Protocol 5 (Gutras)—detailed results regarding the "Gutras" category (comprising the "Brown Gutra" and "White Gutra" classes) from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Gutras)**							
EER	**47.59%**		53.47%		52.62%		55.32%	
HTER	**44.36%**	*41.96%*	53.53%	*46.45%*	53.02%	*45.39%*	52.16%	*46.39%*
Accuracy	**53.57%**	*58.93%*	46.43%	*52.68%*	44.64%	*52.68%*	50.00%	*54.46%*
# True positives	10	*44*	27	*22*	2	*11*	52	*41*
# True negatives	50	*22*	25	*37*	48	*48*	4	*20*
# False positives	3	*31*	28	*16*	5	*5*	49	*33*
# False negatives	49	*15*	32	*37*	57	*48*	7	*18*
Acceptance threshold	0.70	*0.01*	0.97	*0.98*	0.88	*0.83*	0.07	*0.59*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

This specific category includes a collection of both compliant and non-compliant gutras being worn by subjects, therefore the EER is also presented, being calculated from samples with gutras only. The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 21 Protocol 5 (Turban)—detailed results regarding the "Turban" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Turban)**							
HTER	5.88%	*0.00%*	**2.94%**	*0.00%*	26.47%	*0.00%*	29.41%	*0.00%*
Accuracy	88.24%	*100.0%*	**94.12%**	*100.0%*	47.06%	*100.0%*	41.18%	*100.0%*
# True positives	15	*17*	16	*17*	8	*17*	7	*17*
# True negatives	0	*0*	0	*0*	0	*0*	0	*0*
# False positives	0	*0*	0	*0*	0	*0*	0	*0*
# False negatives	2	*0*	1	*0*	9	*0*	10	*0*
Acceptance threshold	0.70	*0.06*	0.97	*0.95*	0.88	*0.71*	0.07	*0.00*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Table 22 Protocol 5 (Others)—detailed results regarding the "Others" category from Table 12

Method	**Ours**		**ICAONet**		**Ours**		**ICAONet**	
Train	**DFIC-HC**				**AR+FRGC**			
Test	**DFIC-HC (Others)**							
EER	**3.85%**		43.24%		71.53%		57.69%	
HTER	**4.07%**	*3.45%*	43.67%	*39.74%*	71.96%	*50.00%*	47.48%	*44.42%*
Accuracy	**95.80%**	*96.64%*	66.39%	*63.87%*	39.50%	*78.15%*	40.34%	*47.90%*
# True positives	25	*25*	10	*14*	2	*0*	19	*18*
# True negatives	89	*90*	69	*62*	45	*93*	29	*39*
# False positives	4	*3*	24	*31*	48	*0*	64	*54*
# False negatives	1	*1*	16	*12*	24	*26*	7	*8*
Acceptance threshold	0.70	*0.87*	0.97	*0.96*	0.88	*0.99*	0.07	*0.19*

Bold values indicate the best result among all compared methods for each metric (i.e., the lowest error rates —EER/HTER — and the highest accuracy)

This specific category includes a collection of both types of head coverings, i.e., religious and not religious, therefore the EER is also presented, being calculated from samples with other head coverings only. The values in italic represent the best possible value for that class, while the ones in black were computed using the optimal general threshold defined in Table 5

Abbreviations

ICAO	International Civil Aviation Organization
ISO	International Organization for Standardization
IEC	International Electrotechnical Commission
MRTD	Machine Readable Travel Document
HC	Head covering
DFIC-HC	Diverse Face Images Coimbra—Head Coverings
YOLO	You Only Look Once
NIST	National Institute of Standards and Technology
C	Compliant
NC	Non-compliant
EER	Equal error rate
HTER	Half total error rate
FRR	False rejection rate
FAR	False acceptance rate
TP	True positives
TN	True negatives
FP	False positives
FN	False negatives
DET	Detection error trade-off

Acknowledgements
We would like to acknowledge all the people that voluntarily took part in the DFIC-HC dataset. We want to acknowledge also all the annotators for their contribution on the manual annotation of the images collected. At last but not the least, we must thank as well to all the people that helped collecting the images, from gathering participants to taking the pictures. In particular, we would like to acknowledge our colleague Ajnas Muhammed, who conducted a big campaign in India that allowed to greatly increase the images of the Asian group on the DFIC-HC dataset.

Author contributions
All authors contributed equally to the paper, except for implementation, which was coded by the two first authors.

Funding
The authors would like to thank the Portuguese Mint and Official Printing Office (INCM) and the Institute of Systems and Robotics-University of Coimbra for the support of the project FACING2. This work has been partially supported by Fundação para a Ciência e a Tecnologia (FCT) under the project UIDB/00048/2020.

Data availability
The DFIC-HC dataset created will be made available until the date of publication of this manuscript, but for non-commercial research use only. The train and test partitions of DFIC-HC will also be made available, as well as our trained version of ICAONet, to facilitate the reproducibility of the presented results. The ones interested in DFIC-HC data must contact the responsible data manager through email (icao.visteam@isr.uc.pt) and agree to the Terms on the License Agreement and transfer agreement documents defined for this transfer. After confirmation, a temporary link and a temporary password to access the server hosting the dataset (images and annotations), will be shared to the requester institution. The public repository is in this link https://github.com/visteam-isr-uc/DFIC-HC.

Declarations

Ethics approval and consent to participants
The dataset was submitted to the Ethics Committee of the University of Coimbra as well as all the legal documents used throughout the process of collection of the DFIC-HC images: ● Declaration of consent for data processing; ● Privacy policy and processing of personal data; ● License agreement for non-commercial research use of DFIC-HC dataset (for the future, the dataset will be made available to the research community). The dataset was approved under the term CEIUC_11/R-9_20241211.

Consent for publication
The image acquisition for the DFIC-HC dataset was done under the consent of all participants that duly signed a Consent Term authorizing their images to be used in a dataset for scientific purposes. Some of them also authorized their images to be included in scientific publications as this paper. The Consent Terms and Privacy Policy showed and explained to each participant were carefully revised by a Law Office under the principles of the General Data Protection Regulation (GDPR). In the case of children, their parents or legal guardian signed the Consent Terms.

Competing interests
Not applicable.

Received: 6 June 2024 Accepted: 14 July 2026
Published online: 29 July 2026

References

1. ISO/IEC 19794-5:2011. ISO/IEC 19794-5:2011 Information technology—Biometric data interchange formats—Part 5: Face image data (2011)
2. ISO/IEC. ISO/IEC 29794-5:2025 Information technology—Biometric sample quality—Part 5: Face image data (2025)
3. ICAO Doc 9303 - Machine Readable Travel Documents - Part 3: Specifications Common to all MRTDs - Eighth Edition (2021)
4. A. Wolf, ICAO: Portrait Quality (Reference Facial Images for MRTD), Version 1.0. Standard. International Civil Aviation Organization (2018),
5. ICAO Portrait Quality (Reference Facial Images for MRTD), (2018)
6. D. Maltoni, A. Franco, M. Ferrara, D. Maio, A. Nardelli, Biolab-icao: A new benchmark to evaluate applications assessing face image compliance to iso/iec 19794-5 standard. 2009 16th IEEE International Conference on Image Processing (ICIP) (pp. 41–44) (2009)
7. A. Martinez, R. Benavente, The AR face database. Tech. Rep. 24 CVC Technical Report (1998,01)
8. P. Phillips, P. Flynn, T. Scruggs, K. Bowyer, J. Chang, K. Hoffman, W. Worek, Overview of the face recognition grand challenge. 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) (Vol. 1, p. 947-954 vol. 1) (2005)
9. A. Kasinski, A. Florek, A. Schmidt, The PUT face database. Image Processing and Communications (Vol. 13.3-4, p. 59-64) (2008)
10. N. Domenico, G. Borghi, A. Franco, D. Maltoni, ONOT: a High-Quality ICAO-compliant Synthetic Mugshot Dataset, in Proceedings of the 18th IEEE International Conference on Automatic Face and Gesture Recognition. (2024)
11. M. Ferrara, A. Franco, D. Maio, D. Maltoni, Face image conformance to iso/icao standards in machine readable travel documents. IEEE Trans. Inf. Forensics Secur. **7**(4), 1204–1213 (2012)
12. BioLab (2000). FVC-ongoing BioLab Benchmark. https://biolab.csr.unibo.it/fvcongoing/UI/FormBenchmarkAreas/BenchmarkAreaFICV.aspx9. (Accessed: 2023-06-30)
13. A. Silva, H. Gomes, L. Batista, A collaborative deep multitask learning network for face image compliance to ISO/IEC 19794–5 standard. Expert Syst. Appl. **198**, 116756 (2022)
14. Federal Office for Information Security. Open Source Face Image Quality (OFIQ) – Draft Report. https://www.bsi.bund.de/SharedDocs/Downloads/EN/BSI/OFIQ/Projektabschlussbericht_OFIQ_1_0.pdf?__blob=publicationFile&v=6. (Accessed: 2024-08-27) (2024)
15. J. Hernandez-Ortega, J. Fierrez, L.F. Gomez, A. Morales, J.G. de Suso, F. Zamora-Martinez. FaceQvec: Vector quality assessment for face biometrics based on ISO compliance. Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (pp. 84–92) (2022)
16. C. Guerra, J. Marcos, N. Gonçalves, Automatic validation of icao compliance regarding head coverings: an inclusive approach concerning religious circumstances, in *2023 international conference of the biometrics special interest group (biosig)*. (2023), pp.1–4
17. ISO/IEC 19794-5:2005. ISO/IEC 19794-5:2005 Information technology—Biometric data interchange formats—Part 5: Face image data (2005)
18. ISO/IEC 39794-5:2019. ISO/IEC 39794-5:2019 Information technology—Extensible biometric data interchange formats—Part 5: Face image data (2019)
19. ISO/IEC. ISO/IEC TR 29794-5:2010 Information technology—Biometric sample quality—Part 5: Face image data (2010)
20. ISO/IEC 2382-37:2022. ISO/IEC 2382-37:2022 - Information technology—Vocabulary, Part 37: Biometrics (2022)
21. C.-H. Lee, Z. Liu, L. Wu, P. Luo, Maskgan: Towards diverse and interactive facial image manipulation, in *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*. (2020), pp. 5549–5558
22. C. Yan, L. Meng, L. Li, J. Zhang, Z. Wang, J. Yin, B. Zheng, Age-invariant face recognition by multi-feature fusionand decomposition with self-attention. ACM Trans. Multim. Comput. Commun. Appl. **18**(1s), 1–18 (2022)
23. National Institute of Standards and Technology (2024). Face Technology Evaluations - FRTE/FATE. https://www.nist.gov/programs-projects/face-technology-evaluations-frtefate. (Accessed: 2024-05-24)
24. J. Redmon, S. Divvala, R. Girshick, A. Farhadi, You only look once: Unified, real-time object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. (2016), pp. 779–788
25. Roboflow (2023). Roboflow: Whats New In YOLOv8. https://blog.roboflow.com/whats-new-in-yolov8/. (Accessed: 2023-06-30)
26. C.-Y. Wang, A. Bochkovskiy, H.-Y. Liao, YOLOv7: Trainable bag-of-freebies sets new state-of-the-art for real-time object detectors, in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. (2023), pp. 7464–7475
27. K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016). https://doi.org/10.1109/LSP.2016.2603342
28. I. Loshchilov, F. Hutter. Decoupled weight decay regularization. arXiv preprint arXiv:1711.05101 (2017),

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.